

A Data Parallel Strategy for Aligning Multiple Biological Sequences on Homogeneous Multiprocessor Platform

Xiangyuan Zhu^{1,2}, Kenli Li², Renfa Li²

¹The Education Technology and Computer Center

Zhaoqing University

Zhaoqing, China

hnzxy@hnu.edu.cn

²College of Information Science and Engineering

Hunan University

Changsha, China

jt_lkl@hnu.cn, lirenfa@vip.sina.co

Abstract—In this paper, we address the biological sequence alignment problem, which is a fundamental operation performed in computational biology. We employ the data parallelism paradigm that is suitable for handling large-scale processing to achieve a high degree of parallelism. Using data parallelism, we propose a strategy in which we employ a parallel clustering scheme to partition the set of sequences into subsets based on sequence similarity. Then the subsets are distributed among the processors using a heuristic algorithm based on Integer Programming so as to minimize the overall processing time, and each subset can be independently aligned in parallel using any sequential approach. The global alignment is achieved using a progressive profile-profile alignment within and between the processors. We implement the proposed algorithm on a cluster using the MPI library, and analyze the experimental results for different problem sizes in terms of quality of alignment, execution time and speed-up.

Keywords—multiple sequence alignment; parallel algorithm; high performance computing

I. INTRODUCTION

Multiple Sequences Alignment(MSA) is a problem of paramount importance and is a fundamental operation performed in computational biology. It provides a wealth of information related to the evolutionary relationships, identifies conserved motifs, and improves structure prediction for RNA and proteins.

With the rapid development of large-scale sequencing techniques, more and more complete

genome sequences are now available in public databases and the number of sequences is rapidly increasing. On the other hand, MSA is a problem of combination in nature. Possibly introducing different number of spaces(gaps) into different positions of sequences, different alignment results are obtained. For two sequences with x number of residues each, there are as much as $(1+\sqrt{2})^{2x+1}\sqrt{x}$ possible alignment combinations[1]. MSA has been proved to be a NP-complete problem. Consequently, there is an urgent need to develop effective and scalable strategies for the computationally intensive operation and vast amount of available data. Parallel solution is a promising approach and various parallel algorithms have been developed in order to speed up the alignment procedure[2],[3].

One of the first exact methods in the literature to locally align two sequences is Smith-Waterman(SW)[4], which is modified from Needleman-Wunsch(NW) algorithm[5]. It is based on dynamic programming and calculates a similarity matrix of size $m \times n$, where m and n are the sizes of the sequences. SW has time and space complexity $O(mn)$. In the SW algorithm, most of the time is spent on calculating the similarity matrix D and this is the part which is usually parallelized. Over the years many parallel variations of the basic SW algorithm are proposed[6],[7].

Because of computational dependency of the similarity matrix elements, a parallelization strategy known as the wavefront method is used since the calculations that can be done in parallel evolve as waves on diagonals[8],[9]. In this method, the maximum parallelism is attained only in the diagonal of the matrix. At other times, one or more processors

are idle, which leads to lower system utilization.

In 2007, a parallel SW algorithm on mesh-based multiprocessor architectures is proposed in [10]. It divides the similarity matrix into three matrixes which store the data of diagonal, row and column respectively. In order to reduce the space complexity of the SW algorithm, a parallel strategy based on distributed shared memory is proposed in [11] in 2008. The space complexity reduces to $O(n)$ through transforming the original bi-dimensional similarity matrix into two linear arrays. The time complexity remains $O(mn)$. In 2009, a divide-and-conquer approach based on SW algorithm is proposed in [12]. It considers two case of sequence alignment where trace-back processes may or may not be done at individual processors. All these strategies mentioned above fundamentally exploit the underlying parallelism embedded in the computational steps of multiple sequence algorithms rather than implementing the data parallelism. Despite the usefulness of these underlying parallelism, they scale very poorly with an increasing number of sequences.

Sample-Align-D, a domain decomposition strategy, is proposed in [13] in 2009. Despite exploiting data parallelism and achieving a super-linear speedups on multiprocessors, this approach has high computation cost of $O(\frac{N}{p})^4 + O(\frac{N}{p})L^2$, where N is the number of sequences, L is the average length of sample sequences, and p is the number of processors.

In this paper, we propose and evaluate Cluster-Distribute-Align, a novel data parallel strategy for aligning multiple sequences on homogeneous multiprocessor platforms. For that we combine and adapt the algorithms proposed by [14], and propose a heuristic data distribution strategy based on Integer Linear Programming.

II. PROBLEM FORMULATION

A. Multiple Sequence Alignment

We define the MSA problem formally. Suppose a biological sequence consists of l characters taken from an alphabet Σ . For DNA sequences, Σ contains 4 different characters, i.e. A, C, G, T. For protein

sequences, Σ contains 20 different amino acids. All the characters in Σ are called residues. Let $S=(s_1, s_2, \dots, s_n)$ be a set of n sequences, where $s_i=s_{i1}s_{i2}\dots s_{il_i}$ ($1 \leq i \leq n$) and l_i is the length of the i th sequence. The solution to MSA problem can be represented as a matrix $A = (a_{ij})$, where $1 \leq i \leq n$, $1 \leq j \leq l$, and $\max(l_i) \leq l \leq \sum_{i=1}^n l_i$. The matrix A has three characteristics.

(1) $a_{ij} \in \Sigma \cup \{\}$, where “.” denotes space(or gap). (2)

If spaces have been deleted, each row of A , i.e. $a_i=a_{i1}a_{i2}\dots a_{il_i}$ ($1 \leq i \leq n$), is the same as corresponding sequence s_i . (3) There is no column which only consists of spaces.

B. Analysis of Parallel Strategy

A data partition and distribution strategy is not only crucial for the performance and scalability of a parallel algorithm, but also for load balancing. An ideal parallel strategy for MSA should satisfy the following requirements.

1) Load balancing: In homogeneous parallel system, all computers have the same processing (computing and I/O) performance. The completion time of sequential MSA on each processor is proportional to its load. Therefore, the problem of load balancing is to look for a partition of a given set of sequences into a given number of subsets such that the loads, i.e. the sums of completion time of MSA in the subsets, are equal or at least nearly equal.

2) Minimized communication time: The subsets should be placed such that the communication cost is minimized. To achieve this goal, each concurrent process of MSA should avoid accessing those subsets located on any of the other processors, because the access of the remote data requires some form of communication.

3) Minimized processing time: The objective of a parallel strategy is to minimize the overall processing time which is proportional to the length of the sequences and the number of sequences. Therefore, the subsets should be placed such that the makespan (we will provide a formal definition in the section III) is minimized. Moreover, a reduction of processing time will amount to a reflection of load balancing.

متن کامل مقاله

دریافت فوری ←

ISIArticles

مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات